结构化数据的隐私与数据效用度量模型 *

谢明明 [a, b]，彭长根 [b, c†]，吴睿雪 [c,d]，丁红发 [a,b]，刘波涛 [b,c]

(贵州大学 a.数学与统计学院; b.贵州省公共大数据重点实验室; c.计算机科学与技术学院; d.密码学与数据安全研究所，贵阳 550025)

摘　要：针对隐私保护中数据隐私量和数据效用的量化问题，基于度量空间和范数基本原理提出了一种结构化数据隐私与数据效用度量模型。首先，给出数据数值化处理方法，将数据表转变为矩阵进行运算；其次，引入隐私偏好函数，度量敏感属性随时间的变化；然后，分析隐私保护模型，量化隐私保护技术产生的变化；最后，构建度量空间，给出了隐私量、数据效用和隐私保护程度计算式。通过实例分析，所建立的度量模型能够有效反映隐私信息量。

关键词：隐私保护；隐私度量；度量空间；隐私量；数据效用

中图分类号：TP309.2　　doi: 10.19734/j.issn.1001-3695.2018.10.0833

Privacy and data utility metric model for structured data

Xie Mingming[a,b], Peng Changgen[b,c†], Wu Ruixue[c,d], Ding Hongfa[a,b], Liu Botao[b, c]

(a. *College of Mathematics & Statistics, b. Guizhou Province Key Laboratory of Public Big Data, c.College of Computer Science & Technology, d. Institute of Cryptography & Data Security, Guizhou University, Guiyang* 550025, *China*)

Abstract: Aiming at the quantification of data privacy and data utility in privacy protection, based on the basic principles of metric space and norm, this paper proposed a privacy and data utility metric model. First, it gave the data numerical processing method. The data was converted into a matrix for calculation. Secondly, it introduced a privacy preference function to measure the change of sensitive attributes over time. Then, it analyzed the privacy protection model and quantified the data changes generated by the privacy protection technology. Finally, this paper built a metric space, and gave privacy amount, data utility and privacy protection calculations. Simulation experiments show that the established metric model can effectively reflect the amount of private information.

Key words: privacy protection; privacy metric; metric space ; privacy amount; data utility

0　引言

如今已经进入了数据的时代，数据渗透在每一个行业和业务职能领域，成为重要的生产要素。在现实生活中，有很多机构的数据需要定期对外发布，比如：医疗数据，交通数据，政务数据等等。这些数据存在着大量的个人隐私信息，一旦泄露将会带来不可估量的损失。在数据发布领域，为了防止隐私数据完全对外公开，数据发布机构通常采取一定的隐私保护技术手段隐藏用户的敏感属性。处理后的数据是否还会泄露隐私，隐私量有多大，对数据可用性造成多大的影响，这些因素是影响数据发布的关键因素。如若不能有效度量隐私及数据效用，将会面临有数据不敢发布的困境，从而导致数据资源开放共享程度低、数据价值难以被有效挖掘利用，因此隐私度量的研究迫在眉睫。

隐私度量方法分为三类，一是根据概率统计方法，利用概率分布信息来推理推断隐私信息的可能性来度量隐私泄露风险；二是利用信息熵[7]，根据信息系统中隐私信息的不确定度来度量隐私信息，三是结合集对分析理论[17]，是一种定性定量相结合并能解决确定与不确定性问题的方法。

针对概率统计方法的隐私度量方法，在 2007 年和 2010 年，Li 等人[1-2]基于 k-匿名和 l-多样性提出了一种计算敏感属性值分布的度量方法，引用 EMD(earth mover’s distance)方法计算数据中敏感属性值的全局分布和任意等价类中同一敏感属性值分布的差异度，差异度越小，隐私信息泄露风险越小。由于 EMD 方法没有考虑等价类与数据间敏感属性值分布的稳定性，在 2014 年，Zhang 等人[3]基于 EMD 方法和 KL 散度提出了一种 EKM 度量方法，通过分布差异度和稳定性差异度两层因素来度量隐私泄露风险大小。根据敏感属性值的概率分布，在 2015 年-2017 年，文献[4~6]提出了一种基于贝叶斯推理的度量隐私信息泄露的方法，通过分析比较推测的信息与隐私信息之间的差异度来度量隐私信息泄露的风险，两者之间的差异度越小，隐私信息泄露风险越大。

信息熵[7](information entropy)作为通信理论的基础，是一种量化信息不确定性的方法。针对利用信息熵的隐私度量方法，在 2002 年，Díaz 等人[8]最早将信息熵应用于隐私保护，提出用信息熵来度量匿名通信系统的匿名性。2006 年，Claußß 等人[9]引用信息熵描述数据集中隐私信息的不确定性。2007 年，Hoh 等人[10]基于信息熵来度量轨迹跟踪的不确定

收稿日期：2018-10-12；**修回日期：**2018-12-28　　**基金项目：**国家自然科学基金资助项目（61662009，61772008）；国家“十三五”密码发展基金资助项目（MMJJ20170129）；贵州省科技计划资助项目（黔科合基础［2016］2315，黔科合基础［2017］1045，黔科合重大专项字［2017］3002，黔科合重大专项字［2018］3001）

作者简介：谢明明（1993-），男，湖北荆州人，硕士研究生，主要研究方向为隐私保护与大数据安全；彭长根（1963-），男（通信作者），贵州贵阳人，教授，博导，博士，主要研究方向为隐私保护、密码学与大数据安全（peng_stud@163.com）；吴睿雪（1995-），女，四川西昌人，硕士研究生，主要研究方向为隐私保护与大数据安全；丁红发（1988-），男，贵州贵阳人，讲师，博士研究生，主要研究方向为分组密码、访问控制及大数据安全；刘波涛（1991-），男，湖南株洲人，硕士研究生，主要研究方向为主要研究方向为分组密码与大数据安全.

度，提出了一个新的时间混淆度量来表示匿名的位置轨迹的隐私程度。2009 年，Ma 等人[11, 12]采用信息理论方法，将隐私量化为位置信息与特定的个人联系的不确定性来量化每个用户的位置隐私水平。Shokri 等人[13]提出了一种基于扭曲的隐私度量方法，通过比较攻击者观察得到的跟踪用户的运动轨迹与用户真实运动轨迹之间的差异来反映用户的隐私水平。2011 年，Chen 等人[14]提出运用条件熵来度量 LBS 中的查询隐私程度，用以测量用户在 LBS 中的查询隐私。2012 年，Yang 等人[15]从网络访问的敏感信息识别个人身份入手，提出了两种类型的攻击者，运用熵度量这两种类型的攻击对一般的网络用户的威胁。2016 年，彭等人[16]为了使信息熵的度量更为直观，将隐私保护系统描述成为一种通信模型，提出了几种隐私保护信息熵模型，从理论的角度上给出了具有通用特性的隐私度量方法。

集对分析理论[17]是具有一定联系的两个集合之间的互相关系、制约、影响的集合对子，通过建立同、异、反联系数从而刻画事物共有属性的确定与不确定关系。在 2015 年，Yan 等人[18]提出一种新的用户隐私保护度量集对分析方法，在数据库隐私保护、位置隐私保护和轨迹隐私保护三种不同应用模式下，建立了隐私度量的体系标准和内容。

文献[19~23,28~30]也对隐私度量的研究进行了相关描述和研究，由此看来，对隐私信息度量的方法不断在发展，并且使用的方法理论也不断更新，考虑更为全面的隐私度量方法有待深入研究。本文为了使隐私信息量表达的更为直观，从数学的角度，借助于度量空间的基本理论，提出一种结构化数据的隐私与数据效用的度量模型。

1　相关理论

1.1　度量空间

度量空间，在数学中是指一个集合，并且该集合中的任意元素之间的距离是可定义的。

定义 1　度量空间。设 R 是一个非空的集合，R 中的元素称为点，对 R 中任意两个点 x,y 都给定一个实数 $\rho(x,y)$ 与他们对应，并且满足：

a) $\rho(x,y)\ge 0$，$\rho(x,y)=0$ 当且仅当 $x=y$；

b)对任意的点 $z\in R$，$\rho(x,y)\le\rho(x,z)+\rho(y,z)$.

称 $\rho(x,y)$ 是两点 x,y 之间的距离，称 R 按照距离 $\rho(x,y)$ 成为度量空间，记为 (R,ρ) 。

根据度量空间的定义，可以得出如下性质：

c)　$\rho(x,y)=\rho(y,x)$ （对称性）

d)　对任何 $x,y,z\in R$，$|\rho(x,z)-\rho(y,z)|\le\rho(x,y)$.

1.2　向量和矩阵范数

范数是泛函分析中的一个基本概念，常常被用来度量某个空间中每个元素的长度或大小。下面介绍向量空间和矩阵空间的范数。

设 $\boldsymbol{n}=(n_1,n_2,\cdots,n_k)$ 是一个向量，$A=(a_{i,j})_{m\times n}$ 是一个矩阵：

向量 1-范数 $\|\boldsymbol{n}\|_1=\sum_{i=1}^{k}|n_i|$；

向量 2-范数 $\|\boldsymbol{n}\|_2=\left(\sum_{i=1}^{k}n_i^2\right)^{\frac{1}{2}}$；

矩阵 F-范数 $\|A\|_F=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}a_{i,j}^2\right)^{\frac{1}{2}}$.

1.3　差分隐私

差分隐私[25]是一种基于数据失真的隐私保护技术，即对敏感数据添加随机噪声使得敏感数据失真从而达到隐私保护的目的，同时保持一定的数据效用。差分隐私的出发点是通过添加随机噪声，使得两个至多相差一条记录的数据集不可区分，避免通过查询结果来推断个体的隐私信息。

定义 2　差分隐私。设数据集 D 和 D' 具有相同的属性结构，两个数据集至多相差一条数据记录，M 为随机算法，$Range(M)$ 为算法 M 的取值范围，$O\subset Range(M)$ 是数据集上的输出结果，如果算法 M 满足

$$\frac{\Pr(M(D)\in O)}{\Pr(M(D')\in O)}\le e^{\varepsilon}$$

称算法 M 满足 ε-差分隐私保护，称 ε 为隐私保护预算。通过限制 ε 的大小来控制隐私保护程度，即 ε 越小，添加的随机噪声越大，隐私保护程度越高，但数据效用越低；同理，ε 越大，添加的随机噪声越小，隐私保护程度越低，数据效用越高。

针对数值型数据，可以通过添加拉普拉斯分布(*Laplace*)的噪声来提供 ε-差分隐私保护。

设随机变量 x 的概率密度函数为

$$f(x\mid\mu,b)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$$

其中：μ 是位置参数，b 是尺度参数，且 $b\ge 0$，称随机变量 x 服从参数 μ,b 的拉普拉斯分布，即 $x\sim Laplace(\mu,b)$ 。它的累计分布函数为

$$F(x)=\int_{-\infty}^{x}f(t)dt=\begin{cases}\frac{1}{2}e^{-\frac{(x-\mu)}{b}}, & x<\mu\\ 1-\frac{1}{2}e^{-\frac{(x-\mu)}{b}}, & x\ge\mu\end{cases}=\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(x-\mu)\left(1-e^{-\frac{|x-\mu|}{b}}\right)$$

它的逆累计分布函数为

$$F^{-1}(p)=\mu-b\operatorname{sgn}(p-\frac{1}{2})\ln(1-2\left|p-\frac{1}{2}\right|) .$$

通过服从均匀分布的随机数和拉布拉斯分布的逆累计分布函数来产生服从拉普拉斯分布的随机数，从而对数据添加噪声，以满足差分隐私保护。

1.4　变量、符号及名词相关说明

1)名词解释

a)数据效用是指经过处理之后的数据与没有处理的同组数据的相同程度或者真实程度，数据真实性越高，数据效用越好；

b)隐私偏好时效性是指同一个体对同一敏感属性的重视程度会随时间的增加而改变。比如，个体患有疾病“肿瘤”，在患病期间，他并不希望别人知道他所患的疾病，此时，他对疾病这一敏感数据重视程度高；但在他康复之后，他会认为让别人知道他曾经所患的“肿瘤”疾病对现在没有影响，因此，随着时间的迁移，该个体对疾病这一敏感属性重视程度越来越低。

2)变量及符号说明

本文涉及到的矩阵的相关符号及运算说明，设 A 是一个 m 行 n 列矩阵

$$A=\begin{bmatrix}a_{1,1} & a_{1,2} & \cdots & a_{1,n}\\ a_{2,1} & a_{2,2} & \cdots & a_{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m,1} & a_{m,2} & \cdots & a_{m,n}\end{bmatrix}$$

矩阵 A 简写为 $A=(a_{i,j})_{m\times n}$，又设 B 是一个与矩阵 A 行和列相同的矩阵 $B=(b_{i,j})_{m\times n}$，相关运算表示如下：

$$+: A+B=(a_{i,j}+b_{i,j})_{m\times n};$$

$\bullet : k\bullet A=(k\bullet a_{i,j})_{m\times n}$，$k$ 是一个实数；

$\otimes : A\otimes B=(a_{i,j}\bullet b_{i,j})_{m\times n}$；

$\odot : A\odot B=(c_{i,j})_{m\times n}, c_{i,j}=\begin{cases} a_{i,j}/b_{i,j}, & (b_{i,j}\neq 0) \\ 1, & (b_{i,j}=0) \end{cases}$；

$\max : \max(A)=\max\limits_{i}\max\limits_{j}\{a_{i,j}\}$ 。

2　隐私与数据效用度量模型

为了度量数据发布中结构化数据的隐私量和经过隐私保护技术处理后的数据效用，本文基于泛函分析中度量空间基本原理构建一个适用于结构化数据的隐私与数据效用度量模型。首先，给出结构化数据进行数值化处理方法，将结构化数据进行数值化得到敏感数据矩阵；其次，考虑用户隐私受主观因素的影响，结合隐私偏好的时效性，引入三类隐私偏好的函数，用以描述敏感数据的敏感性变化过程；然后，分析隐私保护模型对敏感数据矩阵带来的变化，并在敏感数据矩阵的基础上进行量化；最后，构建敏感数据矩阵之间的距离，用以度量隐私和数据的效用。

2.1　数据数值化处理

在结构化数据中，每一个体记录的属性可以分为四类，即：显示标志符属性(explicit identifier attribute)、准标志符属性(quasi-identifier attribute)、敏感属性(sensitive attribute, SA)和非敏感属性(nonsensitive attribute, NA)。由于显示标志符属性在数据发布中会直接去除，准标志属性一般具有数据数值化方法，非敏感属性不在隐私保护的范围内，因此，本文只对敏感属性进行数值化处理。一般情况下，本文将含敏感属性的结构化数据表按表 1 的形式进行描述。

表 1　含敏感属性的结构化数据形式化表

Table 1　Formal table of structured data with sensitive attributes

	SA_1	SA_2	$\cdots$	SA_n
D_1	$data_{1,1}$	$data_{1,2}$	$\cdots$	$data_{1,n}$
D_2	$data_{2,1}$	$data_{2,2}$	$\cdots$	$data_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
D_m	$data_{m,1}$	$data_{m,2}$	$\cdots$	$data_{m,n}$

其中 D_i 表示第 i 个个体（用户），SA_j 表示第 j 个敏感属性，$data_{i,j}$ 表示第 i 个个体的第 j 个敏感属性值。

定义 3　非负数值映射。设 X 是一个含有限个非数值元素的集合，f 是一个映射，如果对每个 $x\in X$，满足 $f(x)\in\mathbb{R}^{+}\cup\{0\}$，则 f 为非负数值映射，所有的非负数值映射构成的集合记为 F 。

根据每一个敏感属性的自身敏感性的特点，按照数据敏感性越敏感映射数值越大的原则，选取 n 个非负数值映射 $f_1, f_2,\cdots, f_n$，将表 1 进行数据数值化计算，即

$$d_{i,j}=f_j(data_{i,j}), i=1,2,\cdots,m,\ j=1,2,\cdots,n$$

得到结构化数据数值化处理结果，用 $D=(d_{i,j})_{m\times n}$ 表示为敏感数据矩阵。

2.2　隐私偏好量化

一般而言，在结构化数据表中，敏感属性中的敏感度是按敏感信息泄露之后所造成的影响进行等级划分。比如，个体所患疾病这一栏敏感属性，“肿瘤”的敏感度要比“感冒”的敏感度高。但实际情况中，个体看待自身数据是否为敏感数据是一个模糊的概念，个人隐私偏好表现了用户对自己隐私数据不愿被披露的程度。

定义 4　隐私偏好向量。设 p_i 是某一个体对敏感属性 SA_i 的数据不愿被披露程度权重值，由同一个体对每一敏感属性数据披露程度权重值组成的向量 $\boldsymbol{p}=(p_1, p_2,\cdots, p_n)$ 满足

$$\|\boldsymbol{p}\|_1=1, 0\le p_i\le 1(i=1,2,\cdots,n)$$

称 $\boldsymbol{p}$ 为这一个体的隐私偏好向量。

个体的隐私偏好向量可以由个体的主观评价确定，也可以通过个体的历史数据分析推断得出。个体的隐私偏好向量的 2-范数 $\|\boldsymbol{p}\|_2$ 反映了个体的隐私偏好类型：$\|\boldsymbol{p}\|_2$ 值接近 1 时，表现了个体越集中重视某一或多个敏感属性，而忽略其他敏感属性；$\|\boldsymbol{p}\|_2$ 值接近 $\sqrt{1/n}$ 时，表现了个体不具有隐私偏好，对每一敏感属性重视程度相同。

设 $\boldsymbol{p}_i=(p_{i,1}, p_{i,2},\cdots, p_{i,n})$ 为第 i 个个体的隐私偏好向量，由所有个体的隐私偏好向量构成的矩阵记为隐私偏好矩阵 P，即：$P=(\boldsymbol{p}_1, \boldsymbol{p}_2,\cdots, \boldsymbol{p}_m)^{\mathrm{T}}=(p_{i,j})_{m\times n}$ 。通过对敏感数据矩阵 D 与 P 的合成 $G=D\otimes P=(d_{i,j}\bullet p_{i,j})_{m\times n}$ 得到带有隐私偏好的敏感数据矩阵 G 。

进一步，为了描述隐私偏好随时间迁移的变化情况，量化隐私偏好的时效性，本文引入三类隐私偏好函数。

定义 5　隐私偏好函数。设 $\varphi(t)$ 是 $[0,+\infty]$ 上的函数，$a,b\in[0,1]$，且 $a<b$，满足

$$\begin{cases} \inf\limits_{t\in[0,+\infty]}\{\varphi(t)\}=a \\ \sup\limits_{t\in[0,+\infty]}\{\varphi(t)\}=b \end{cases}$$

称 $\varphi(t)$ 为隐私偏好函数。

第一类隐私偏好函数。个体看待某一敏感属性会随时间的增加，表现出越来越重视该敏感属性，且重视程度有上限，即 $\varphi(t)$ 为有界递增函数；

第二类隐私偏好函数。个体看待某一敏感属性会随时间的增加，表现出越来越忽视该敏感属性，且忽视程度有下限，即 $\varphi(t)$ 为有界递减函数；

第三类隐私偏好函数。个体对待某一敏感属性的重视程度，不会随时间的增加而受到影响，即 $\varphi(t)$ 为常数函数。

这三类函数能够简明地描述个体对敏感属性的偏好随时间增加的变化情况，而其他复杂的类型均可由这三类隐私偏好函数分段构成。类似于隐私偏好矩阵，本文构建隐私偏好函数矩阵。设个体 D_i 的敏感属性 SA_j 的隐私偏好函数为 $p_{i,j}(t)$，由全部个体的每一敏感属性的偏好函数组成的矩阵 $P(t)=(p_{i,j}(t))_{m\times n}$ 称为隐私偏好函数矩阵。同样，将敏感数据矩阵 D 与 $P(t)$ 合成 $G(t)=D\otimes P(t)=(d_{i,j}\bullet p_{i,j}(t))_{m\times n}$ 得到带有隐私偏好函数的敏感数据矩阵 $G(t)$ 。

2.3　隐私保护模型分析

在数据发布前，为了保护数据中的隐私信息，需要使用隐私保护技术对数据进行处理，在数据发布领域隐私保护技术可以分为两大类：基于加密的隐私保护技术和基于非加密的隐私保护技术。本文将对结构化数据使用隐私保护技术视为将敏感数据矩阵 D 进行相应的变化，具体如下：

a)基于加密的隐私保护技术是隐私保护效果最好的一类方法，加密的数据不会暴露任何隐私信息，但加密的数据会造成数据不可用。因此，使用基于加密的隐私保护技术后，敏感数据矩阵变为零矩阵，即：$D\to D'=\boldsymbol{0}$；

b)基于非加密的隐私保护技术又可分为基于数据失真的隐私保护技术和基于数据匿名的隐私保护技术。数据失真方法会造成数据结果与真实数据发生一定的偏差，本文将这一偏差视为敏感数据矩阵上的偏差，也就是使用基于数据失真的隐私保护技术后，敏感数据矩阵变为数据加噪后的矩阵，即：$D\to D'=D+\Delta D$，其中 $\Delta D=(\Delta d_{i,j})_{m\times n}$，$\Delta d_{i,j}$ 为 $d_{i,j}$ 的偏差。数据匿名方法一定程度上能够使攻击者不能判别隐私信息所

属的具体个体，从而保护了个人隐私。数据匿名的基本原理是让发布的数据中存在一定数量的不可区分的数据，但攻击者可以以一定的概率得到原始的敏感数据，从敏感数据矩阵角度来看，使用基于数据匿名的隐私保护技术后，会以一定概率隐藏敏感信息，敏感数据矩阵 D 中的元素 $d_{i,j}$ 会以一定概率 $q_{i,j}$ 变成 0，$d_{i,j}\xrightarrow{q_{i,j}}0$，$q_{i,j}\in[0,1]$，对于敏感数据矩阵产生的变化，本文使用期望来量化，即 $D\to D'=((1-q_{i,j})\bullet d_{i,j})_{m\times n}$。

通过的隐私保护模型的量化分析，可以衡量隐私保护技术的隐私保护程度和使用隐私保护技术后数据的效用。

2.4　隐私与数据效用度量

通过结构化数据的数值化、隐私偏好的量化和隐私保护模型的分析，将对结构化数据中隐私和数据效用的量化分析，转换为对敏感数据矩阵的度量，能够更直观地了解数据中的隐私信息。

结构化数据数值化的敏感数据矩阵所构成的集合记为 $\mathcal{D}$，$D_1,D_2\in\mathcal{D}$，满足 $\|D_1\|_F=\|D_2\|_F$ 时，称 $D_1=D_2$（这里的“=”并非两个矩阵相同），设距离

$$\rho(D_1,D_2)=|\|D_1\|_F-\|D_2\|_F|$$

则集合 $\mathcal{D}$ 与距离 ρ 构成度量空间。

证明　集合 $\mathcal{D}$ 为非空集合，对任意的 $D_1,D_2\in\mathcal{D}$，有 $\rho(D_1,D_2)=|\|D_1\|_F-\|D_2\|_F|\geq 0$，并且

$$\rho(D_1,D_2)=0\Leftrightarrow|\|D_1\|_F-\|D_2\|_F|\Leftrightarrow\|D_1\|_F=\|D_2\|_F\Leftrightarrow D_1=D_2$$

满足定义 1 的性质 a)；对任意的 $D_3\in\mathcal{D}$，

$$\begin{aligned}\rho(D_1,D_3)+\rho(D_2,D_3)&=|\|D_1\|_F-\|D_3\|_F|+|\|D_2\|_F-\|D_3\|_F|\\&\geq|\|D_1\|_F-\|D_3\|_F+\|D_3\|_F-\|D_2\|_F|\\&=|\|D_1\|_F-\|D_2\|_F|\\&=\rho(D_1,D_2)\end{aligned}$$

满足定义 1 的性质 b)，因此构成 $(\mathcal{D},\rho)$ 度量空间。

结构化数据的所含的隐私信息量转换为敏感数据矩阵的隐私量来衡量，敏感数据矩阵是度量空间 $(\mathcal{D},\rho)$ 的点，从而定义度量空间 $(\mathcal{D},\rho)$ 点的大小，这里自然采用范数来定义。由于每个结构化数据表敏感属性特点不同，非负数值映射的值域不同，因此在计算敏感数据矩阵的隐私量时，将数据归一化。

定义 6　隐私量。设 $D\in\mathcal{D}$，用 $|D|$ 表示敏感数据矩阵的隐私量，则

$$|D|=\left\|\frac{D}{\max\{D\}}\right\|_F=\frac{1}{\max\{D\}}\|D\|_F$$

根据定义 6，带有隐私偏好的敏感数据矩阵 G 的隐私量为 $|G|=\|G/\max\{G\}\|_F$，带有隐私偏好函数的敏感数据矩阵 $G(t)$ 的隐私量为 $|G(t)|=\|G(t)/\max\{G(t)\}\|_F$。

数据效用的度量需要有一个参考点，由于所研究的数据发布没有指定的发布环境，不知道接收方需要何种数据，因此，本文采用与原始数据相比较的方法来度量丢失的信息，即：将原始数据表中每一个数据的数据量记为 1，经过处理后数据的数据量取值范围为 $[0,1]$。

定义 7　数据效用。设 D 是原始敏感数据矩阵，D' 是 D 经过处理过的敏感数据矩阵，且具有相同的结构，D 和 D' 的数据量分别用 $U(D)$ 和 $U(D')$ 来表示，则

$$U(D)=\|D\odot D\|_F,U(D')=\|D'\odot D\|_F$$

称 $U(D'|D)=U(D')/U(D)$ 为数据 D 经过处理后的数据效用。$U(D'|D)$ 的取值范围为 $[0,1]$，其值越接近 1，表明处理后的数据越接近真实值。

对隐私保护程度的度量也是隐私度量的一个重要方面，它与数据效用一起可以评价隐私保护模型的好坏。目前，越来越多的研究者在不断改进隐私保护模型，达到隐私保护程度高并且数据效用也高的目的。本文采用使用隐私保护技术处理后的数据隐私量的减少情况来衡量隐私保护程度。

设 M 是一个隐私保护算法，它将敏感数据矩阵 D 变成 D'，算法 M 对 D 的隐私保护程度

$$L_M(D)=\frac{\rho(\max\{D'\}\bullet D,\max\{D\}\bullet D')}{\|\max\{D'\}\bullet D\|_F}.$$

2.5　模型适用范围

文献[19]较为全面地列出了隐私度量方法，如不确定性度量、信息增加或损失度量、数据集相似性度量、不可区分性度量、敌手攻击成功概率度量和时间、误差和精度的度量等。本文所提出的结构化数据的隐私与效用度量模型是针对于数据发布领域中结构化数据的隐私信息量、数据可用性和隐私保护程度的量化，其方法可以用于信息损失的度量、数据集相似性度量，但不限于此，如：用户身份认证过程中用户身份的匿名性、登录行为的不可追踪性[26, 27]可用基于匿名的隐私保护技术量化方法进行量化，访问控制结构中的访问策略可将其数据结构化后进行度量。本文所提出的隐私度量模型其目的是为了让数据拥有者更为直观地掌握数据中的隐私信息。下面通过实例分析进行展现。

3　实例分析

政务数据包含了大量的个人信息，非常具有挖掘价值，但政务数据包含的敏感数据太多，导致政务部门不敢发布数据，采取的是与数据接收机构签约合作的方式进行数据共享。隐私与数据效用的度量能够使数据发布者有效把握发布数据的隐私和数据的可用性，对评估隐私泄露风险有重要的意义。为了体现实验的准确性和科学性，采用公开的 UCI 机器学习数据库中的 Adult 数据来完成实验分析，该数据包含了 48842 条记录和 14 项属性。下面选取包含年龄(Age)、学历(Education)和职业(Occupation)属性的前 5 条记录作为实验室数据集 D_1，并在 D_1 的基础上做简单的修改作为实验数据集 D_2（表 2），数据集 D_2 用于与数据集 D_1 的隐私量进行对比分析，并用差分隐私对数据集 D_1 进行保护来对度量模型进行分析，分析数据的隐私量、带隐私偏好数据的隐私量、数据效用和隐私保护程度。

表 2　两组实验数据表

Table 2　Two experimental data tables

D_1	age	education	occupation
1	39	Bachelors	Adm-clerical
2	50	Bachelors	Exec-managerial
3	38	HS-grad	Handlers-cleaners
4	53	Bachelors	Handlers-cleaners
5	28	Bachelors	Prof-specialty
D_2	Age	Education	Occupation
1	39	Bachelors	Adm-clerical
2	50	Bachelors	Exec-managerial
3	38	HS-grad	Exec-managerial
4	53	Bachelors	Handlers-cleaners
5	28	Bachelors	Prof-specialty

按照“age”“education”和“occupation”三类敏感数据的敏感级别构建非负数值映射：

收入： $f_1(x)=1-\left|\frac{x-25}{25}\right|, x\in[0,50], f_1(x)=0, x>50$ ；

职业： $f_2: Bachelors\to 0.50$， $HS-grad\to 0.71$ ；

病症： $f_3: \begin{matrix} Adm-clerical\to 0.95 \\ Exec-managerial\to 0.65 \\ Handlers-cleaners\to 0.34 \\ Prof-specialty\to 0.78 \end{matrix}$ ；

经过数据数值化处理得到两组表格敏感数据矩阵结果为

$$D_1=\begin{bmatrix} 0.44 & 0.50 & 0.95 \\ 0.00 & 0.50 & 0.65 \\ 0.48 & 0.71 & 0.34 \\ 0.00 & 0.50 & 0.34 \\ 0.88 & 0.50 & 0.78 \end{bmatrix},\quad D_2=\begin{bmatrix} 0.44 & 0.50 & 0.95 \\ 0.00 & 0.50 & 0.65 \\ 0.48 & 0.71 & 0.65 \\ 0.00 & 0.50 & 0.34 \\ 0.88 & 0.50 & 0.78 \end{bmatrix}$$

根据定义 6 敏感数据隐私量的定义，计算两组数据表的隐私量， $|D_1|=2.3223$ ， $|D_2|=2.3944$ 。从结果可以看出 D_2 的隐私信息量比 D_1 的大，按照“Occupation”属性职业人数越少隐私性越大的原则，该结果与表 2 相符合。

下面给出隐私偏好函数矩阵，分析个人隐私偏好对数据中隐私量的变化过程。设 $P(t)(t\in[0,100])$ 为数据表 D_1 的隐私偏好函数矩阵

$$P(t)=\begin{bmatrix} 0.33 & 0.33 & 0.34 \\ \frac{0.33}{1+0.005t} & \frac{0.33}{1+0.005t} & \frac{0.34+0.005t}{1+0.005t} \\ \frac{0.33+0.004t}{1+0.001t} & \frac{0.33}{1+0.001t} & \frac{0.34-0.003t}{1+0.001t} \\ \frac{0.33+0.003t}{1+0.006t} & \frac{0.33}{1+0.006t} & \frac{0.34+0.003t}{1+0.006t} \\ \frac{0.33}{1+0.006t} & \frac{0.33}{1+0.006t} & \frac{0.34+0.006t}{1+0.006t} \end{bmatrix}$$

带有隐私偏好函数的敏感数据矩阵为 $D_1(t)=D_1\otimes P(t)$ 的隐私量 $|D_1(t)|$ 随时间的变化过程如图 1 所示。

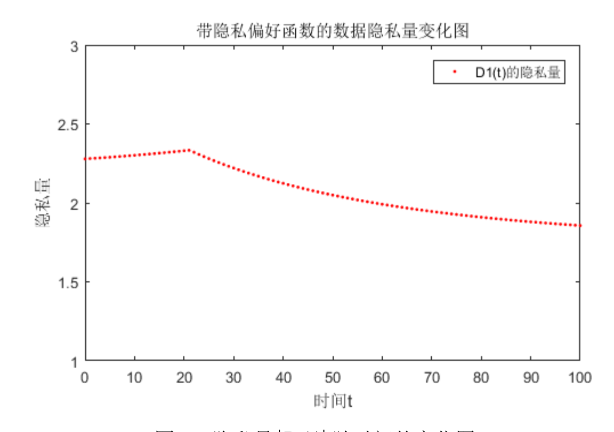


图 1　隐私量 $|D_1(t)|$ 随时间的变化图

Fig. 1　Change of $|D_1(t)|$ over time

从图 1 可以看出带有隐私偏好函数数据集 D_1 的隐私量随时间增加先增加再逐渐减少，并在 1.8 附近趋于稳定。说明该数据发布后，个体对部分敏感数据重视程度降低。

接下来使用差分隐私保护技术 M，分析数据的隐私，数据效用，隐私保护程度的量的变化。

通过拉普拉斯分布的逆累计分布函数和均匀分布随机序列来生成满足拉普拉斯分布的随机数。设 a 是均匀分布 $[-0.5,0.5]$ 上的一个随机数，均值 $\mu=0$ ，标准差为 σ ，则满足拉普拉斯分布的随机数

$$x=\mu-\frac{\delta}{\sqrt{2}}\mathrm{sgn}(a)\bullet\ln(1-2|a|)$$

标准差为 σ 反映了添加噪声的大小， σ 越小添加的噪声就越小。

取 $\sigma=0.01$ ，针对数据集 D_1 生成随机噪声

$$\Delta D_1=\begin{bmatrix} 0.0052 & 0.0296 & 0.0068 \\ 0.0329 & 0.0066 & -0.0024 \\ -0.0070 & -0.0012 & -0.0136 \\ 0.0058 & 0.0043 & 0.0014 \\ -0.0066 & 0.0000 & 0.0121 \end{bmatrix}$$

然后对数据集 D_1 进行添加噪声，当原始数据敏感矩阵值小于噪声值，则加噪后变为 0，

$$D_1'=D_1-|\Delta D_1|=\begin{bmatrix} 0.4348 & 0.4704 & 0.9432 \\ 0.0000 & 0.4934 & 0.6476 \\ 0.4730 & 0.7088 & 0.3264 \\ 0.0000 & 0.4957 & 0.3386 \\ 0.8734 & 0.5000 & 0.7679 \end{bmatrix}$$

根据隐私量、数据效用的定义和隐私保护程度的计算式，计算得到隐私量 $|D_1'|=2.3127$ ，数据效用 $U(D'|D)=0.9877$ ，隐私保护程度 $L_M(D)=0.0041$ ，可以看出通过差分隐私保护后，数据集的隐私量有所下降，由于 σ 的取值很小，所以隐私保护程度低，数据效用高。当逐步增加添加的噪声时，隐私保护程度会变高，数据效用变低，变化过程如图 2 所示。

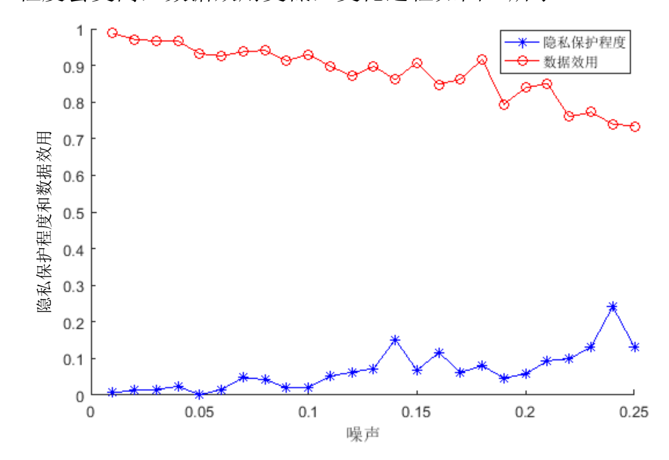


图 2　隐私保护程度与数据效用随噪声增加的变化图

Fig. 2　The change of degree of privacy protection and data utility with noise

通过实例分析所构建的隐私与数据效用度量模型中的隐私量、隐私偏好、数据效用和隐私保护程度可以得出，所建立的模型能够有效反映出数据中的隐私量以及隐私保护技术处理后的隐私保护程度和数据效用，为数据发布者从定量的角度有效把握数据中的隐私，从而为评估数据发布泄露风险提供一定的依据。

4　结束语

隐私度量方法目前还没有一套完善的理论，目前常用的基于概率统计、基于信息论和基于集对分析理论的或多或少存在一定的缺陷。本文提出了一种结构化数据的隐私与数据效用度量模型，试图从构建度量空间的角度出发，建立信息之间的距离来衡量信息的差异，并对信息本身的隐私量大小进行的定义。为了度量数据发布中结构化数据的隐私量和经过隐私保护技术处理后的数据效用，首先，由于非数值型数据不可计算，因此给出结构化数据进行数值化处理方法，将结构化数据进行数值化，转为对数据矩阵的计算；其次，用户隐私受主观因素的影响，结合隐私偏好的时效性，引入三类隐私偏好的函数，用以简明描述一般情况下隐私的敏感性

变化过程；然后，分析隐私保护模型的特点，对隐私保护技术的改变进行量化；最后，构建敏感数据矩阵之间的距离，用以度量隐私和数据的效用。经过实例分析，本文所提出的模型能够有效反映数据的隐私量和数据效用的变化，也可以作为隐私保护程度和数据效用两者之间博弈的量化方法。

参考文献：

[1] Li Ninghui, Li Tiancheng, Venkatasubramanian S. *t*-closeness: privacy beyond *k*-anonymity and *l*-diversity [C]//Proc of the 23rd International Conference on Data Engineering. Piscataway,NJ: IEEE Press, 2007: 106-115.

[2] Li Ninghui, Li Tiancheng, Venkatasubramanian S. Closeness: a new privacy measure for data publishing [J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22 (7): 943-956.

[3] Zhang Jianpei, Xie Jing, Yang Jing, *et al.* A *t*-closeness privacy model based on sensitive attribute values semantics bucketization [J]. Journal of Computer Research and Development, 2014, 51 (1): 126-137.

[4] Gkountouna O,Terrovitis M. Anonymizing collections of tree-structured data [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27(8): 2034-2048.

[5] Yuji Y, Kouichi I. k-presence-secrecy: practical privacy model as extension of *k*-anonymity [J]. IEICE Trans. on Information& System, 2017 (4): 730-740.

[6] Li Xiangyang, Zhang Chunhong, Jung T, *et al.* Graph-based privacy-preserving data publication [C]// Proc of the 35th Annual IEEE International Conference on Computer Communications. Piscataway,NJ: IEEE Press, 2016: 1-9.

[7] Shannon C E. A mathematical theory of communication [J]. Bell System Technical Journal, 1948, 27(3): 379-423.

[8] Díaz C, Seys S, Claessens J, *et al.* Towards measuring anonymity [C]// Proc of the 2nd International Conference on Privacy Enhancing Technologies. Berlin: Springer, 2002: 54-68.

[9] Clauß S, Stefan S. Structuring anonymity metrics [C]// Proc of the 2nd ACM Workshop on Digital Identity Management. New York: ACM Press, 2006: 55-62.

[10] Hoh B, Gruteser M, Xiong Hui, *et al.* Preserving privacy in gps traces via uncertainty-aware path cloaking [C]// Proc of the 14th ACM conference on Computer and communications security. New York: ACM Press, 2007: 161-171.

[11] Ma Zhendong, Kargl K, Weber M. A location privacy metric for V2X communication systems [C]// Proc of IEEE Sarnoff Symposium. Piscataway,NJ: IEEE Press, 2009: 1-6.

[12] Ma Zhendong, Kargl K, Weber M. Measuring location privacy in V2X communication systems with accumulated information [C]// Proc of the 6th IEEE International Conference on Mobile Adhoc and Sensor Systems. Piscataway,NJ: IEEE Press, 2009: 322-331.

[13] Shokri R, Freudiger J, Jadliwala M, *et al.* A distortion-based metric for location privacy [C]// Proc of the 8th ACM Workshop on Privacy in the Electronic Society. New York: ACM Press, 2009: 21-30.

[14] Chen Xihui, Pang Jun. Measuring query privacy in location-based services [C]// Proc of the 2nd ACM Conference on Data and Application Security and Privacy. New York: ACM Press, 2012: 49-60.

[15] Yang Yuhao, Lutes J, Li Fengjun, *et al.* Stalking online: On user privacy in social networks [C]// Proc of the 2nd ACM Conference on Data and Application Security and Privacy. New York: ACM Press, 2012: 37-48.

[16] 彭长根，丁红发，朱义杰，等. 隐私保护的信息熵模型及其度量方法 [J]. 软件学报, 2016, 27(8): 1891-1903. (Peng Changgen, Ding Hongfa, Zhu Yijie, *et al.* Information entropy models and privacy metrics methods for privacy protection [J]. Journal of Software, 2016, 27(8): 1891-1903.)

[17] 赵克勤. 集对分析及其初步应用 [M]. 杭州: 浙江科学技术出版社, 2000. (Zhao Keqin. Set pair analysis and its preliminary application [M]. Hangzhou: Zhejiang Science and Technology Press, 2000.)

[18] 晏燕，郝晓弘，王万军. 一种隐私保护度量的集对分析方法 [J]. 武汉大学学报:工学版，2015, 48(6): 883-890.(Yan Yan, Hao Xiaohong, Wang Wanjun. A set pair analysis method for privacy. metric [J]. Engineering Journal of Wuhan University, 2015, 48 (6): 883-890.)

[19] Wagner I, Eckhoff D. Technical privacy metrics: a systematic survey [J]. ACM Computing Surveys, 2018, 51(3): articleNo 57.

[20] 王玲玲，马春光，刘国柱. 基于位置服务的隐私保护机制度量研究综述 [J]. 计算机应用研究, 2017, 34(3): 647-652. (Wang Lingling, Ma Chunguang, Liu Guozhu. Survey on metrics for location-based privacy protection mechanisms [J]. Application Research of Computers, 2017, 34(3): 647-652.)

[21] 熊金波，王敏燊，田有亮，等. 面向云数据的隐私度量研究进展 [J]. 软件学报, 2018,29(7): 1963-1980. (Xiong Jinbo, Wang Minshen, Tian Youliang, *et al.* Research progress on privacy measurement for cloud data [J]. Journal of Software, 2018,29(7): 1963-1980.)

[22] 王璐，孟小峰. 位置大数据隐私保护研究综述 [J]. 软件学报, 2014, 25(4): 693-712. (Wang Lu, Meng Xiaofeng. Location privacy preservation in big data era: a survey [J]. Journal of Software, 2014, 25(4): 693-712.)

[23] 张学军，桂小林，伍忠东. 位置服务隐私保护研究综述 [J]. 软件学报，2015, 26(9): 2373-2395. (Zhang Xuejun, Gui Xiaolin, Wu Zhongdong. Privacy preservation for location-based services: a survey [J]. Journal of Software, 2015, 26(9): 2373-2395.)

[24] 夏道行. 实变函数论与泛函分析. 下册 [M]. 北京: 高等教育出版社, 2010: 1-107. (Xia Daoxing. Real variable function theory and functional analysis. volume 2 [M]. Beijing: Higher Education Press, 2010: 1-107.)

[25] Dwork C, Roth A. The algorithmic foundations of differential privacy [M]. Boston: Now Publishers Inc. 2014.

[26] Wang Ding, Wang Ping. On the anonymity of two-factor authentication schemes for wireless sensor networks: attacks, principle and solutions [J]. Computer Networks, 2014, 73(11): 41-57.

[27] Wang Ding, Wang Ping. Two birds with one stone: two-factor authentication with security beyond conventional bound [J]. IEEE Trans on Dependable and Secure Computing, 2018, 15(4): 708-722.

[28] Dasgupta B, Mobasheri N, Yero I G. On analyzing and evaluating privacy measures for social networks under active attack [J]. Information Sciences, 2019, 473(1): 87-100.

[29] Ahmad A. , Mukkamala R. A novel information privacy metric [C]// Proc of the 14th International Conference on Information Technology. Cham: Springer, 2018: 221-226.

[30] Zhao Yuchen, Wagner I. POSTER: evaluating privacy metrics for graph anonymization and de-anonymization [C]// Proc of Asia Conference on Computer and Communications Security. New York: ACM Press, 2018: 817-819.